DOCUMENT RESUME

ED 410 291 TM 027 135

AUTHOR	Du, Yi; And Others
TITLE	Raters and Single Prompt-to-Prompt Equating Using the FACETS Model in a Writing Performance Assessment.
PUB DATE	Mar 97
NOTE	24p.; Paper presented at the International Objective Measurement Conference (9th, Chicago, IL, March 21, 1997).
PUB TYPE	Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Elementary Education; *Elementary School Students; *Equated Scores; Grade 5; Grade 7; Interrater Reliability; *Performance Based Assessment; Sample Size; *Scoring; Tables (Data); *Writing Tests
IDENTIFIERS	Calibration; FACETS Computer Program; *FACETS Model; Logits; Minneapolis Public Schools MN; Rasch Model; Writing Prompts

ABSTRACT

The FACETS equating model meets the complex requirements for equating writing performance assessment across both raters and prompts. This study is based on an equating of the 1996 writing performance assessment in the Minneapolis Public Schools (Minnesota). Raters and prompts were equated simultaneously using the FACETS model. About 3,000 fifth graders and 3,000 seventh graders participated in the writing assessment. Three prompts were assessed in each grade, and each student wrote to one of the prompts. About 30 raters were selected from Minneapolis Public Schools teachers to score the papers using a uniform rubric. An extension of the Rasch model to include multiple facets (FACETS model) was used in equating to determine the transformation rules. The four facets were student, item (scoring component), rater, and prompt. Overall results show that the FACETS model calibrates raters, students, topics, and scoring dimensions so that all facets are positioned on a common scale. The scale is in log-odds, or "logit," units that constitute an equal-interval scale with respect to appropriately transformed probabilities of responding in particular categories. The advantages of the FACETS model include: (1) sample independence; (2) calibration invariance; (3) equating more than one facet at the same time; and (4) flexibility in the sample size for examinees and items. (Contains 9 figures, 8 tables, and 13 references.) (SLD)

Raters and Single Prompt-to-Prompt Equating Using the FACETS Model in a Writing Performance Assessment

Yi Du

William L. Brown

Char Rogers

Research, Evaluation and Assessment

Minneapolis Public Schools

Paper Presented in the Ninth International Objective Measurement Conference on

March 21, 1997 in Chicago

In a writing performance assessment, multiple prompts for different genres are usually needed because students are expected to be able to write in different genres. Because of the amount of the time required and the cost of the assessment, each student is usually restricted to responding to one or two prompts. It seems evident that test scores derived from different genres will not generally be equivalent. Even when efforts are made in the test construction process to make different prompts as nearly equivalent as possible. However, these efforts are often not sufficient to ensure test score equivalence across different prompts. Besides prompts, rater severity is another key source of variation that makes student scores unequivalent and non-comparable. Unless each rater scores every student paper, part of each student's score will be dependent on who grades the paper as. Therefore, test equating is often used to adjust test scores so that the scores on different forms or prompts, and from different raters, are more nearly equivalent.

A variety of equating models, such as raw score linear equating and equipercentile equating, were considered and have been tried in this study. However, these equating models were developed for machine-scannable multiple choice assessment and can equate prompts, but not raters. Both rater and prompt are primary sources of variation making student scores incomparable. Therefore, it is not appropriate to apply these models to writing assessment.

The FACETS equating model meets the complex requirement for equating writing performance assessment across both raters and prompts. The FACETS model "can provide a framework for obtaining objective and fair measurements of writing ability that are statistically invariant over raters, writing tasks, and other aspects of the writing assessment process." (Engelhard, 1992, p173).

This study is based on an equating of the 1996 writing performance assessment in Minneapolis Public Schools (MPS). In this assessment, raters and prompts were

equated simultaneously using the FACETS model. By presenting the results based on the 1996 assessment, this study attempts to address two issues: First, reliable results of equating both rater and prompt can be obtained using the FACETS model scores. Second, single prompt-to-prompt equating is feasible if the appropriate design and equating model are selected.

Data

About 3,000 Grade 5 students and 3,000 Grade 7 students participated in this writing assessment. Three prompts, representing narrative, persuasive and informative writing within a common topic, were assessed at grades 5 and 7. Each student wrote to one of the three prompts. Students were assigned randomly to specific prompts. (Because the results are similar, we present only Grade 5 student data in this study.)

About thirty raters were selected from the population of Minneapolis Public Schools teachers. The three prompts were scored during three separate sessions in the following order: narrative, informative, and persuasive. Within each session, raters were trained before they scored papers. For each prompt, a representative sample (about 40%) of all papers was scored by two raters. These papers for double scoring were distributed spirally from rater to rater, i.e., each rater was paired with every other rater at least once. After raters were well trained, they scored double-rated papers first. After finishing the double-rated papers, raters scored single-rated papers. This pattern was consistent for all prompts, ensuring that all raters graded all three genres of papers and every rater was linked with all others across these prompts. Figure 1 shows the linkage among raters when they scored the double-rated papers.

RATER 2 \ RATER 1	A	B	C	D	E	F	G
A							
B	X						
C	X	X					
D	X	X	X				
E	X	X	X	X			
F	X	X	X	X	X		
G	X	X	X	X	X	X	

Figure 1. Linkage of Raters Used in Scoring 40% of Papers

A uniform scoring rubric was used to score the three groups of papers. The scoring rubric includes three domains: Purpose and Voice; Organization and Details; and Conventions of Writing. Under each dimension, multiple features were included in the scoring guide. All the scoring features were rated on a "1 to 4" scale. The framework of the scoring rubric is shown in Table 1.

Table 1
The Framework of the Scoring Rubric

	Domain	Scoring Feature	Scale
1	Purpose and Voice	Purpose	1-4
		Voice	1-4
2	Organization	Main Idea	1-4
		Organization	1-4
		Details	1-4
3	Conventions	Sentence Structure	1-4
		Spelling	1-4
		Punctuation/Capitalization	1-4
		Grammar/Usage	1-4
		Legibility	1-4

An analytic scoring method was used in this assessment to provide detailed information about each student's writing, compared with the District Standards, to improve reporting to teachers, students and parents. The scores in the three domains ("Purpose and Voice," "Organization," and "Conventions") were grouped and averaged, yielding three mean scores on a 1-4 scale. A total raw score was then obtained by adding the three scores together. Generally, the overall raw score is derived from these features according to the following formula:

Raw score = average (Purpose + Voice) + average (Main idea + Organization + +Details) + average (Sentence + Spelling + Punctuation + Grammar + +Legibility)

Given that all these writing features are scored on a scale of 1 to 4, based on this formula the raw score ranges from 3 to 12.

Equating Design

The random-groups design was used in this assessment, in which different prompts were administered to different but randomly equivalent groups of students. Under the random-groups equating design, student groups who take different test prompts are regarded as being sampled from the same population. The population of Grade 5 students was divided into three random groups. One of three different prompts (persuasive, narrative and informative) was administered to each group during the testing period. The common rater group links the three individual student groups. Every rater was paired with all of the other raters at least once. A uniform scoring rubric was used to score all the three prompts. Figure 2 shows the general design of raters, students, scoring features and prompts.

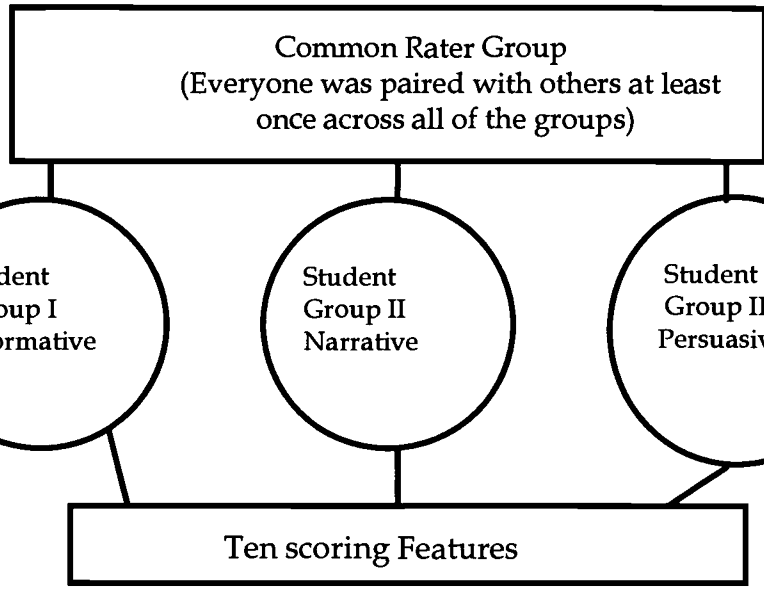

Figure 2. Linkage of Raters, Prompts, Scoring Rubric, and Student Groups

FACETS Model

An extension of the Rasch model to include multiple facets (FACETS model) was used in equating to determine the transformation rules. For the MPS writing assessment, the primary FACETS model includes four facets: student, item (scoring component), rater and prompt:

$$\log\left(\frac{P_{nijmk}}{P_{nijmk-1}}\right) = B_n - D_i - C_j - A_m - F_k \qquad (1)$$

where P_{nijmk} is the probability of student n being graded in category k by rater j on item i and topic m, $P_{nijmk-1}$ is the probability of student n being graded k-1 by rater j on item i and topic m, B_n is the writing ability measure of student n, D_i is the difficulty calibration of item i, C_j is the severity measure of rater j, A_m is the difficulty calibration of prompt k, and F_k is the difficulty calibration of grading category k-1 relative to category K. The rating scale is k=0, K.

Within the FACETS model, the three student groups were anchored to the same group mean. Thus, equating was controlled by the adjustment made for the three student groups based on prompt differences. Because the three equivalent student groups share the same scale with the same group mean and same measurement units, the differences among the prompts can be attributed to the differences of the difficulty level of the prompts and sample errors. Thus, adjustment is made for student measures based on the difficulty of the prompts. Had we not anchored the three groups to the same group mean, students who responded to easier prompts would have appeared to be more able, and students who responded to harder prompts would be appeared to be less able. A variance analysis was conducted to examine the interaction between raters and prompts. The results show that the interaction between raters and prompts is not enough to consider. Therefore, only student groups were anchored in this study.

Prompt Difficulty Equating and Adjustment

As we discussed earlier, student raw scores cannot be assumed to be comparable if they responded to different prompts. Finding that prompts differ substantially in the degree of difficulty can make test developers aware of the prompt differences, and allow them to adjust student scores in accordance with the difficulty of prompts.

The FACETS model produces a measure of the difficulty level of each prompt. Table 2 rank-ordered these prompts from the most difficult at the top to the easiest at the bottom. The informative prompt was hardest, the narrative prompt was easiest, with the persuasive prompt in between. All fit statistics are between 1.0 and 1.1, which indicates that the data from the topics fit the model well enough for measuring student ability. The difficulty differences between the prompts are significant, χ^2 (2) =4997.1 and 2939.5, $p < .001$ with a high separation reliability (R=1.00). This implies

that an equating procedure is necessary to adjust the prompt difficulty for student scores.

Table 2
Prompts Calibration and Analysis

Prompt	Rasch Measure	S.E.	Infit Mean Squares	Outfit Mean squares	Raw Score Average
Informative	0.29	0.01	1.1	1.1	2.4
Narrative	-0.22	0.01	1.1	1.1	2.6
Persuasive	-0.07	0.01	1.0	1.0	2.5
Overall	0	0.01	1.1	1.1	

Figures 3 through 5 show the differences in difficulties of prompts and how the FACETS equating adjusted these differences. In Figure 3, three ogive curves represent the three student groups who produced informative, narrative and persuasive writings, respectively. The conversion between raw scores and the Rasch measures indicates that raw score is dependent on the prompts. Students with the same writing ability receive unfair higher raw scores on narrative writing and unfair lower scores on persuasive and informative writing because of the difficulty of the prompts. After equating, the FACETS model adjusted the difficulty of the prompts for student measures. Thus, student measures for different groups are equivalent and comparable. One may notice that there is little difference between students with greater than 6 logits on the Rasch scale. That may imply that the 1-4 scale has a ceiling effect so that the scale cannot differentiate top students very well. Another possibility could be that these high achieving students are able to write very well to any of the three prompts. Exploration of these possibilities is beyond the scope of this study.

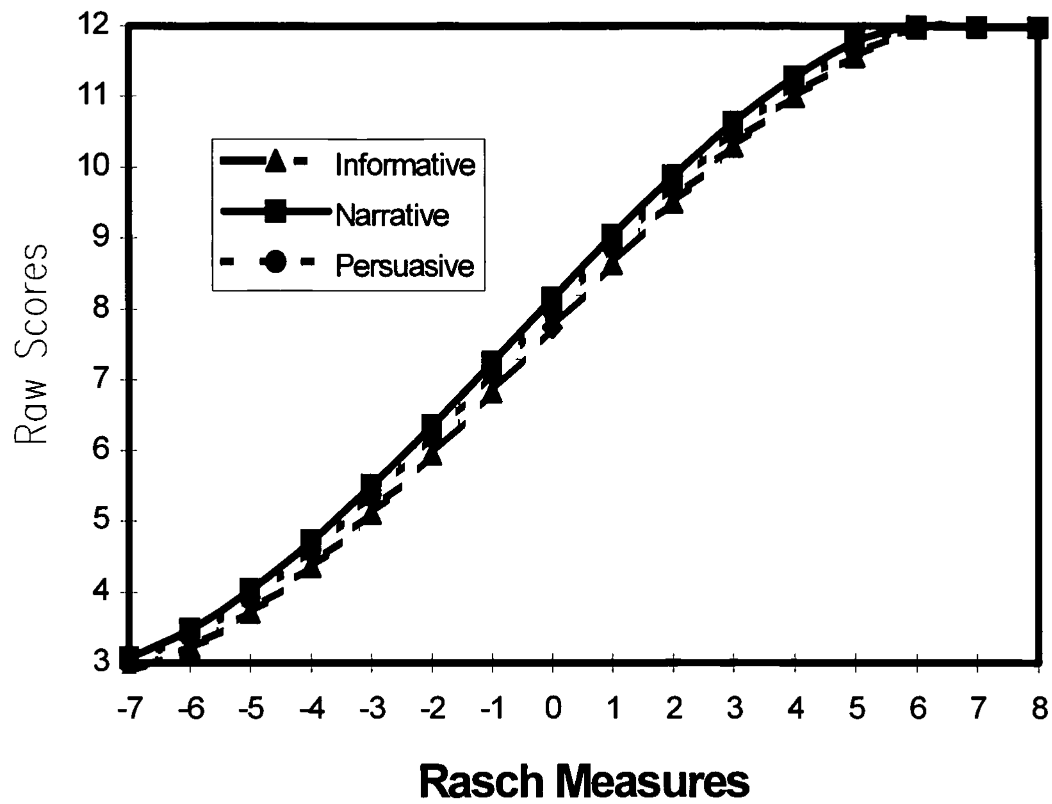

Figure 3. FACETS Equated Measures and Raw Scores on Three Prompts

In order to make the Rasch measures more easily understood by teachers, parents and students, the Rasch measures were transformed linearly to scale ranging from 3 to 12. The new reporting scale looks like, but is quite different from the raw score scale. The reporting scale keeps the good properties of the Rasch scale: prompt difference adjusted, calibration invariance, and equal interval, so that student scores are accurate and comparable. Figure 4 shows the linear relationship between student raw scores and their Rasch measures.

Figure 4. Rasch Measures and Transformed Reporting Scores

Figure 5 shows the relationship between the adjusted reporting scores and the raw scores. This figure indicates how the reporting scale adjusts for students' scores based on prompt difficulties. For example, a student with a raw score of 8 receives a reporting score about 7.9 on narrative writing, 8.0 on persuasive writing, and 8.1 on informative writing. The reporting score makes student results from different prompts comparable.

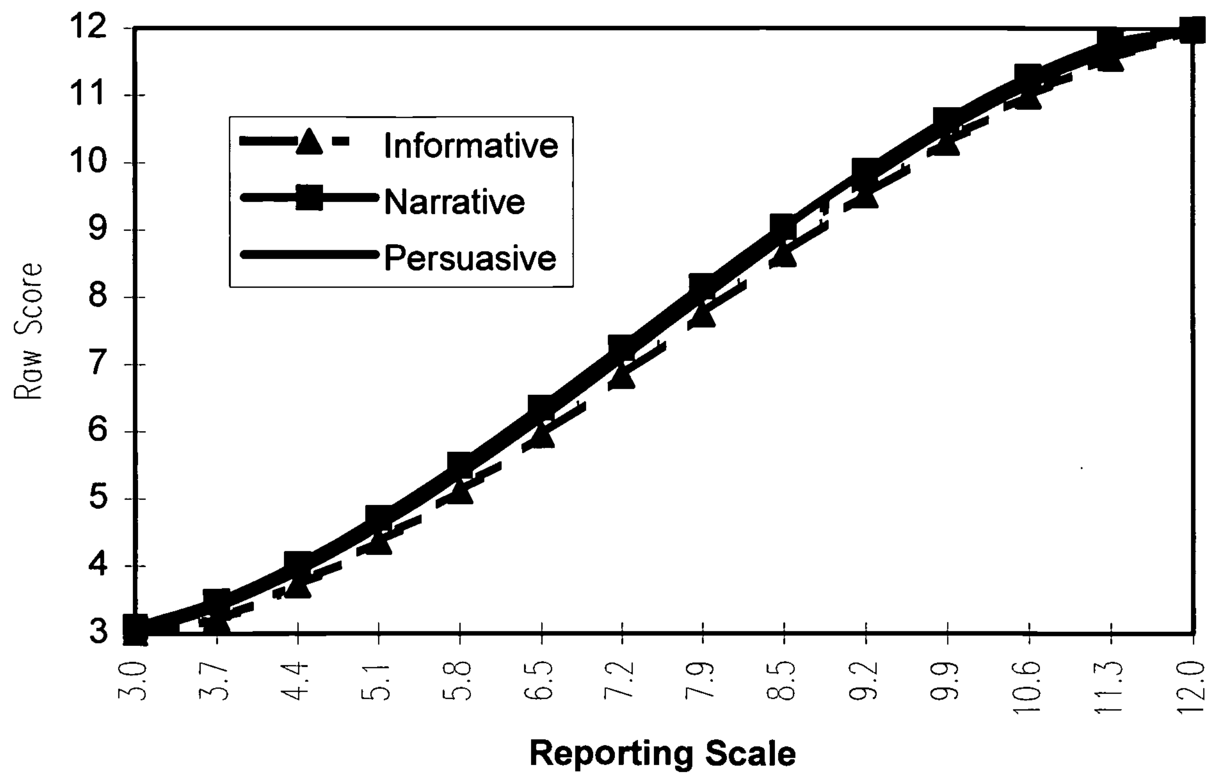

Figure 5. Rasch Reporting Scores and Raw Scores

It will be clear to identify prompt variations in raw scores and adjust through equating if we control the rater variation. Table 3 exhibits student pairs who wrote to different prompts but were rated by the same raters. This table shows how prompt difficulties affect raw scores and how the FACETS equating removes prompt difficulty differences from student measures.

Students "394540" and "835015" were graded by the same raters, and earned the same raw scores on narrative and informative prompts. However, their ability measures are -3.47 and -2.94 logits respectively. The substantial difference of .51 logits occurred because the difficulties of the prompts are different (0.51 logits different). The raw score of the first student (narrative writing) was overestimated because of the easier prompt; the second student (informative writing) was

underestimated because of the harder prompt. The student measures, which are corrected for differences in prompt difficulties, provide fair assessment for the two students.

The other pairs of student measures demonstrate similar patterns. These results show that the raw scores were affected by different prompts and that the FACETS equating process adjusts for student measures based upon prompt difficulties.

Table 3 Prompt Equated and Adjusted on Rasch Scale (Same Raters)

Student	Prompt	Calibration	Raw Score	Rasch Measure
394540	Narrative	-0.22	4.7	-3.47
835015	Informative	0.29	4.7	-2.94
	Diff.	0.51		0.51
075329	Informative	0.29	5.4	-2.22
798274	Narrative	-0.22	5.4	-2.72
	Diff.	0.51		0.50
073933	Persuasive	-0.07	6.5	-1.43
591471	Narrative	-0.22	6.5	-1.58
	Diff.	0.15		0.15
047130	Narrative	0.29	5.5	-2.31
791185	Persuasive	-0.07	5.5	-2.67
	Diff.	0.36		0.36
012067	Persuasive	-0.07	11.8	5.65
799301	Informative	0.29	11.8	6.04
	Diff.	0.36		0.39
598687	Persuasive	-0.07	11.6	4.85
791208	Narrative	-0.22	11.6	4.70
	Diff.	0.15		0.15
081213	Narrative	-0.22	8.9	0.57
397206	Informative	0.29	8.9	1.06
	Diff.	0.51		0.49

Note: Standard Errors for all the prompts are 0.01.

Table 4 uses the reporting scale score to compare with the raw score, instead of the Rasch measures. This table shows that the reporting scores follow the same pattern as the Rasch measures and that the reporting score removes prompt difficulty differences from student scores.

Table 4. Prompt Equated and Adjusted on Rasch Scale Scores (Same Raters)

Student	Prompt	Calibration	Raw Score	Scale Score
394540	Narrative	7.7	4.7	5.5
835015	Informative	8.1	4.7	5.9
	Difference	0.4		0.4
075329	Informative	8.1	5.4	6.3
798274	Narrative	7.7	5.4	5.9
	Difference	0.4		0.4
073933	Persuasive	7.8	6.5	6.9
591471	Narrative	7.7	6.5	6.8
	Diff.	0.1		0.1
047130	Informative	8.1	5.5	6.3
791185	Persuasive	7.8	5.5	6.0
	Diff.	0.3		0.3
012067	Persuasive	7.8	11.8	11.7
799301	Informative	8.1	11.8	12.0
	Diff.	0.3		0.3
598687	Persuasive	7.8	11.6	11.2
791208	Narrative	7.7	11.6	11.1
	Diff.	0.1		0.1
081213	Narrative	7.7	8.9	8.3
397206	Informative	8.1	8.9	8.7
	Diff.	0.4		0.4

Table 5 shows the comparison of group distributions before and after equating. The results indicate that for the different student groups, the means, standard deviations, spreads, and shapes of distributions are equivalent and comparable after equating. Without equating, students have very differing probabilities of success when they write to different prompts.

Table 5. Comparison between Raw Scores and Scale Scores

	Raw Score (Before Equating)			Scale Score (After Equating)		
	Informative	Narrative	Persuasive	Informative	Narrative	Persuasive
N Count	1365	986	969	1365	986	969
Mean	7.8	8.1	8.0	7.9	7.9	7.9
S.D.	2.09	1.09	1.97	1.76	1.65	1.66
Kurtosis	0.08	-0.25	0.30	0.14	0.49	0.52
Skewness	-0.53	-0.13	-0.30	0.28	0.19	0.16

Rater Equating and Adjustment

As we know, student raw scores may not be comparable if they happened to be rated by severe raters. Examining discrepant ratings may not be an appropriate or adequate method for resolving this issue. Two severe raters may agree in their ratings of a student, but without knowing that the two raters are significantly more severe than other raters, one would have no reason to question these ratings. Finding that raters differ substantially in the degree of severity exercised can suggest a need to address such differences in rater training, or to consider the feasibility of adjusting students' scores in accordance with the severity or leniency of the raters.

The FACETS model produces a measure of the degree of severity of each rater. Table 6 (see column labelled "severity measure") rank-orders these raters from the most sever at the top to the most lenient at the bottom. To the right of each Rater Severity Measure is the standard error of the estimate, indicating the precision with which it

has been estimated. Other things being equal, the more observations an estimate is based on, the smaller its standard error. The rater severity ranges from -0.92 to 0.50 at grade 5. The spread is 1.42 logits. This represents a mean score discrepancy of appromixately 0.4 on the 4-point scale. All of the raters are between -1.00 and + 1.00 logit in severity.

Table 6
Rater Severity Analysis

Rater ID	Severity Measure	S.E.	Infit Mean Squares	Outfit Mean squares	Raw Score Average
43	0.46	0.02	1.0	1.0	2.7
37	0.41	0.02	1.2	1.1	2.6
17	0.34	0.02	1.0	1.0	2.5
11	0.28	0.01	1.2	0.7	2.5
14	0.28	0.02	0.7	1.0	2.5
39	0.26	0.02	1.1	0.9	2.7
25	0.2	0.02	0.9	0.9	2.5
20	0.18	0.01	0.9	1.1	2.4
36	0.14	0.02	1.1	1.1	2.6
30	0.1	0.02	1.1	0.9	2.5
33	0.05	0.2	9.0	1.2	2.8
42	0.04	0.02	1.2	0.7	2.8
40	0.01	0.02	0.6	0.9	2.6
35	-0.03	0.01	0.9	1.1	2.7
21	-0.06	0.02	1.1	1.1	2.8
32	-0.13	0.02	1.1	0.9	2.7
34	-0.13	0.02	9.0	1.1	2.7
22	-0.15	0.02	1.1	1.0	2.9
15	-0.16	0.02	1.0	1.1	2.6
19	-0.17	0.02	1.0	1.2	2.7
27	-0.17	0.01	1.0	1.6	2.7
13	-0.18	0.02	1.2	1.1	2.6
18	-0.2	0.02	1.6	1.2	2.6
38	-0.23	0.02	0.9	0.8	2.6
16	-0.28	0.02	1.2	1	2.6
23	-0.3	0.02	1.3	0.9	2.8
12	-0.31	0.02	0.8	1.1	2.7
31	-0.47	0.01	1.1	1.3	2.8
28	-0.49	0.02	1.3	1.4	2.8
26	-0.53	0.02	1.5	1.5	2.8
Overall	0	0.01	1.1	1.1	2.6

Figures 6 through 8 show the raw scores plotted against the Rasch measures within each prompt. These figures illustrate that raw scores unadjusted for rater severity can mask variability in writing competence.

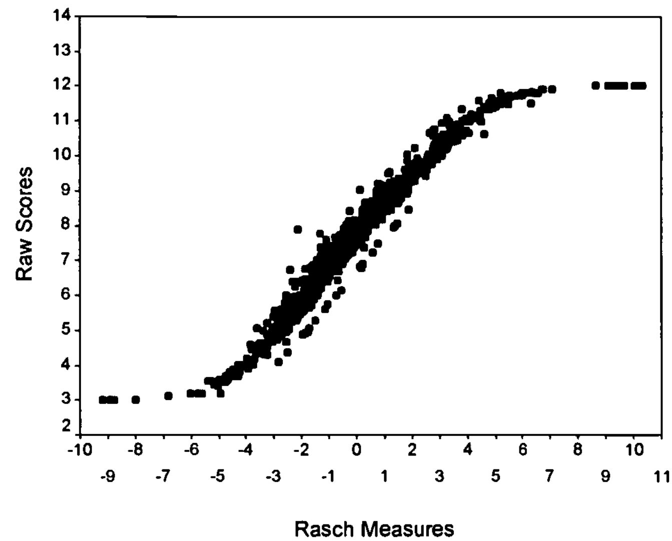

Figure 6 Raw Scores and the Rasch Measures on Informative Writing

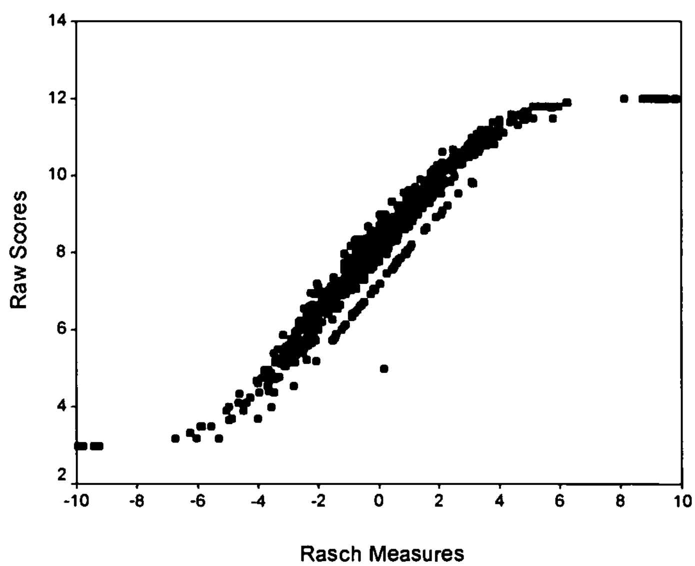

Figure 7 Raw Scores and the Rasch Measures on Narrative Writing

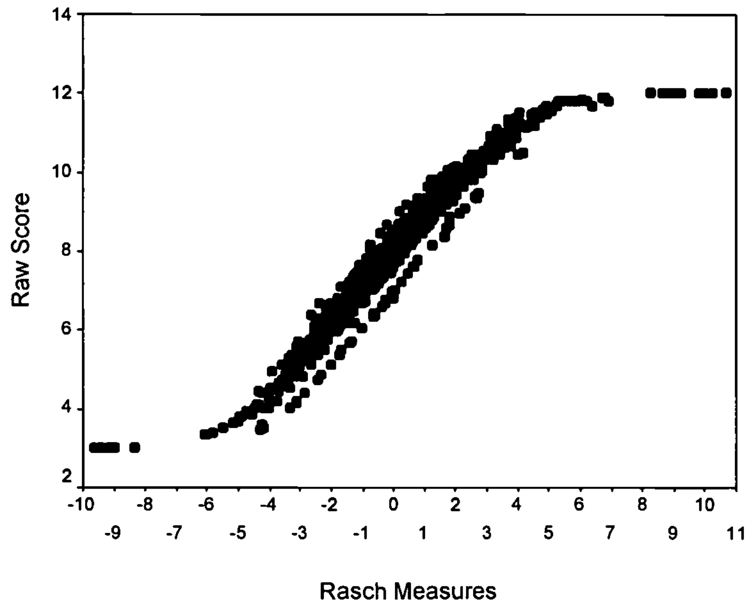

Figure 8 Raw Scores and the Rasch Measures on Persuasive Writing

It is easy to see rater severity differences and adjustment if we control prompt effects. Table 7 shows how rater severity affects raw scores, and how rater severity is removed from student measures when prompt difficulties are controlled. The student pairs in Table 7 wrote to the same prompts, but were graded by different raters. These students were selected for comparison of the measures given by different raters.

Student pair "004107" and "780815" earned the same raw scores graded by different raters, but their ability measures are -0.30 and +0.34 logits respectively. The substantial difference of .64 logits occurred because the different raters have different severity level (0.64 different). Student "004107" had a more severe rater, while Student "780815" had a more lenient rater. The rater severity difference made the two students' raw scores same. The Rasch measures removed the effects of rater severity and provided fair and comparable estimates of writing ability. The same can be said for the other pairs of students.

Table 7 Rater Severity Equated and Adjusted

Student	Rater 1 with Severity	Rater 2 with Severity	Average Severity	Raw Score	Rasch Measure	Prompt
004107	23 (-0.30)		-0.30	7	-1.07	Informative
780815	17 (0.34)		0.34	7	-0.43	Informative
Diff.			0.64		0.54	
691478	16 (-.028)		-.028	4.3	-5.05	Narrative
397613	14 (0.28)		0.28	4.3	-4.49	Narrative
Diff.			0.56		0.56	
793336	17 (0.34)		0.34	13	5.97	Persuasive
592085	28(-0.49)		-0.49	13	5.14	Persuasive
Diff.			0.83		0.83	
012690	33 (0.05)		0.05	11.7	4.88	Narrative
598627	31 (-0.47)		-0.47	11.7	4.37	Narrative
Diff.			0.52		0.51	
781379	32 (-0.13)		-0.13	7.1	-0.83	Informative
080844	14 (0.28)		0.28	7.1	-0.36	Informative
Diff.			0.41		0.47	
243309	23 (-0.30)	24 (-0.73)	-0.52	7.3	-1.01	Informative
591402		25 (0.20)	0.20	7.3	-0.28	Informative
Diff.			0.72		0.73	
399286	43 (0.46)		0.46	8.5	1.25	Informative
595063	18 (-0.20)		-0.20	8.5	0.60	Informative
Diff.			0.66		0.65	
691478	16 (-.028)		-0.28	3.9	-5.05	Narrative
397613	14 (0.28)		0.28	3.9	-4.49	Narrative
Diff.			0.56		0.56	

Overall Results

The overall results for students, raters, prompts and scoring items are shown graphically in Figure 9. The FACETS program calibrates the raters, students, topics and scoring dimensions so that all facets are positioned on a common scale. That scale is in log-odds, or "logit" units which, under the model, constitute an equal-interval scale with respect to appropriately transformed probabilities of responding

in particular categories. The figure enables one to view all facets of the analysis simultaneously, summarizing key information about each facet.

Figure 9 shows that the student distribution spreads from -7 to +8. All raters are located beween +1 logit and -1 logit, which means they are not extremely severe or lenient. The informative prompt was the hardest, while the narrative was the easiest.

```
--------------------------------------------------------------------------------------------------
|Measr|+Student   |-Rater  |-Prompt  |-Item                                                |S.1  |
--------------------------------------------------------------------------------------------------
+   8 + ***       +        +         +                                                     +(4)  +
|     |           |        |         |                                                     |     |
|     |           |        |         |                                                     |     |
+   7 + .         +        +         +                                                     +     +
|     | .         |        |         |                                                     |     |
|     | .         |        |         |                                                     |     |
+   6 + .         +        +         +                                                     +     +
|     | .         |        |         |                                                     |     |
|     | *.        |        |         |                                                     |     |
+   5 + *.        +        +         +                                                     +     +
|     | *.        |        |         |                                                     |     |
|     | .         |        |         |                                                     |     |
+   4 + *.        +        +         +                                                     +     +
|     | **.       |        |         |                                                     |     |
|     | **.       |        |         |                                                     | --- |
+   3 + ***.      +        +         +                                                     +     +
|     | ***.      |        |         |                                                     |     |
|     | ***.      |        |         |                                                     |     |
+   2 + *****     +        +         +                                                     +     +
|     | ****.     |        |         |                                                     | 3   |
|     | *****.    | .      |         |                                                     |     |
+   1 + ******.   + *      +         +                                                     +     +
|     | *******.  |        |         | Gram/Usage   Organization Sentence     Spelling     |     |
|     | ********. | *****  | Inform  | Punc/Capit                                          |     |
*   0 * ********. * *****. * Persuade* Details                                             * --- *
|     | ********. | *****  | Narrate | Purpose                                             |     |
|     | ********. | *      |         | Main Idea                                           |     |
+  -1 + *********.+ .      +         + Legible      Voice                                  +     +
|     | ********. | .      |         |                                                     |     |
|     | ******.   |        |         |                                                     | 2   |
+  -2 + ******.   +        +         +                                                     +     +
|     | *****.    |        |         |                                                     |     |
|     | ***.      |        |         |                                                     |     |
+  -3 + ***.      +        +         +                                                     +     +
|     | **.       |        |         |                                                     | --- |
|     | *.        |        |         |                                                     |     |
+  -4 + .         +        +         +                                                     +     +
|     | *.        |        |         |                                                     |     |
|     | .         |        |         |                                                     |     |
+  -5 + .         +        +         +                                                     +     +
|     | .         |        |         |                                                     |     |
|     | .         |        |         |                                                     |     |
+  -6 + .         +        +         +                                                     +     +
|     | .         |        |         |                                                     |     |
|     | .         |        |         |                                                     |     |
+  -7 + *.        +        +         +                                                     +(1)  +
--------------------------------------------------------------------------------------------------
|Measr| * = 21    | * = 2  |-Prompt  |-Item                                                |S.1  |
--------------------------------------------------------------------------------------------------
```

Figure 9. Maps of Overall Results

This assessment includes four facets: students (about 3000 each grade), raters (about 30, nested with students), prompts (3 prompts nested with students and raters), and scoring items (10 items crossed with raters). Furthermore, 60% of the papers were rated by one rater and 40% by two raters. Rasch-based generalizability can be conducted to estimate the reliability of the assessment when all facets are considered. Table 8 shows the estimated variance analysis for difference facets. The generalizability (or reliablity) estimate is 0.81. The variance components analysis shows that the scoring item facet takes into account the largest variance except the main effect for students. If we want to increase the generalizability, we need to improve the scoring rubric. A variance analysis conducted using a small sample from the population, showed that the magnitudes of interactions between facets were very small (about 0.0001). Therefore, we can assume the variance of the interaction is zero. This table does not include the variances of interaction.

Table 8

Rasch-based Generalizability

Rasch Analysis Results	Student	Rater	Item	Topic
S.D.	2.16	0.5	0.68	0.21
RMSE	0.55	0.09	0.01	0.01
$S.D.^2$= Observed Variance	4.67	0.25	0.46	0.04
$RMSE^2$ = Error Variance	0.30	0.008	0.0001	0.0001
True Variance	4.37	0.24	0.46	0.04
Rasch-based Generalizability	0.81			

Conclusions and Discussion

This study demonstrates the feasibility of using the FACETS model to equate both raters and prompts in a writing performance assessment. It also demonstrates the feasibility of equating prompts. The advantages of the FACETS model--sample independence, calibration invariance, equating more than one facet at the same time, and flexiblity in the sample size for examinees and items--make equating both raters and prompts feasible and ensures accurate and stable results.

REFERENCES

Braun, H. I. (1988). Understanding scoring reliability: Experiments in calibrating essay readers. Journal of Educational Statistics, 13, 1-18.

Burger, S. E. & Burger, D. L. (1994). Determining the validity of performance-based assessment. Educational Measurement: Issues and Practice, 13, 9- 15.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratuam, N. (1972). The dependability of behavioral measurements: Theory of generalizability of scores and profiles. New York: Wiley.

Du, Y. (1995). Bias detection in performance assessment. Paper presented at the Midwest Objective Measurement Seminar, Chicago.

Du, Y. (1995). DIF adjustment. Rasch Measurement SIG Newsletter, 9, 414.

Du, Y. (1995). Measuring writing abilities in a large-scale direct writing assessment. Paper presented at the Eighth International Objective Measurement Workshop, Berkeley: California.

Du, Y., Wright, B. D, & Brown, W. L. (1996). Differential Facet Functioning in Direct Writing Assessment. Paper presented at the Annual Conference of American Educational Research Association, New York.

Engelhard, J. G. (1992a). The measurement of writing ability with a many-faceted Rasch model. Applied Measurement in Education, 5, 171-191.

Engelhard, J. G. (1992b). Historical reviews of invariance: evidence from the measurement theories of Thorndike, Thurstone, and Rasch. Educational and Psychological Measurement, 52, 275-291.

Engelhard, J. G., Gordon, B., & Gabrielson, S. (1992). The influences of model of discourse, experiential demand and gender on quality of student writing. Research in the Teaching of English, 26, 315-336.

Wright, B. D. & Stone, M. H. (1979). Best test design. Chicago: MESA Press.

Wright, B. D. & Masters, G. N. (1982). Rating scale analysis. Chicago: MESA Press.

Wright, B. D., & Linacre M. (1994). Reasonable Mean-square Fit Values. Rasch Measurement SIG Newsletter, 8, p. 370.

RASCH

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

ERIC

TM027135

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Raters and Single Prompt-to-Prompt Equating Using the FACETS Models in a Writing Performance Assessment

Author(s): Yi Du, William L. Brown and Char Rogers

Corporate Source: Minneapolis Public Schools

Publication Date:

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following two options and sign at the bottom of the page.

[X] Check here **For Level 1 Release:** Permitting reproduction in microfiche (4" x 6" film) or other ERIC archival media (e.g., electronic or optical) *and* paper copy.	The sample sticker shown below will be affixed to all Level 1 documents PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) **Level 1**	The sample sticker shown below will be affixed to all Level 2 documents PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) **Level 2**	[ ] Check here **For Level 2 Release:** Permitting reproduction in microfiche (4" x 6" film) or other ERIC archival media (e.g., electronic or optical), but *not* in paper copy.

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at Level 1.

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."

Sign here→ please

Signature: Yi Du	Printed Name/Position/Title: Yi Du / Testing and Assessment Specialist, William L. Brown, Director	
Organization/Address: Minneapolis Public Schools 807 NE Broadway Minneapolis, MN 55413	Telephone: 612/627-2195	FAX: 612/627-2277
	E-Mail Address: yidu@mpls.k12.mn.us	Date: 4-28-97

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC Clearinghouse on Assessment and Evaluation
210 O'Boyle Hall
The Catholic University of America
Washington, DC 20064

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
1100 West Street, 2d Floor
Laurel, Maryland 20707-3598

Telephone: 301-497-4080
Toll Free: 800-799-3742
FAX: 301-953-0263
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com

(Rev. 6/96)